A EXAMPLES

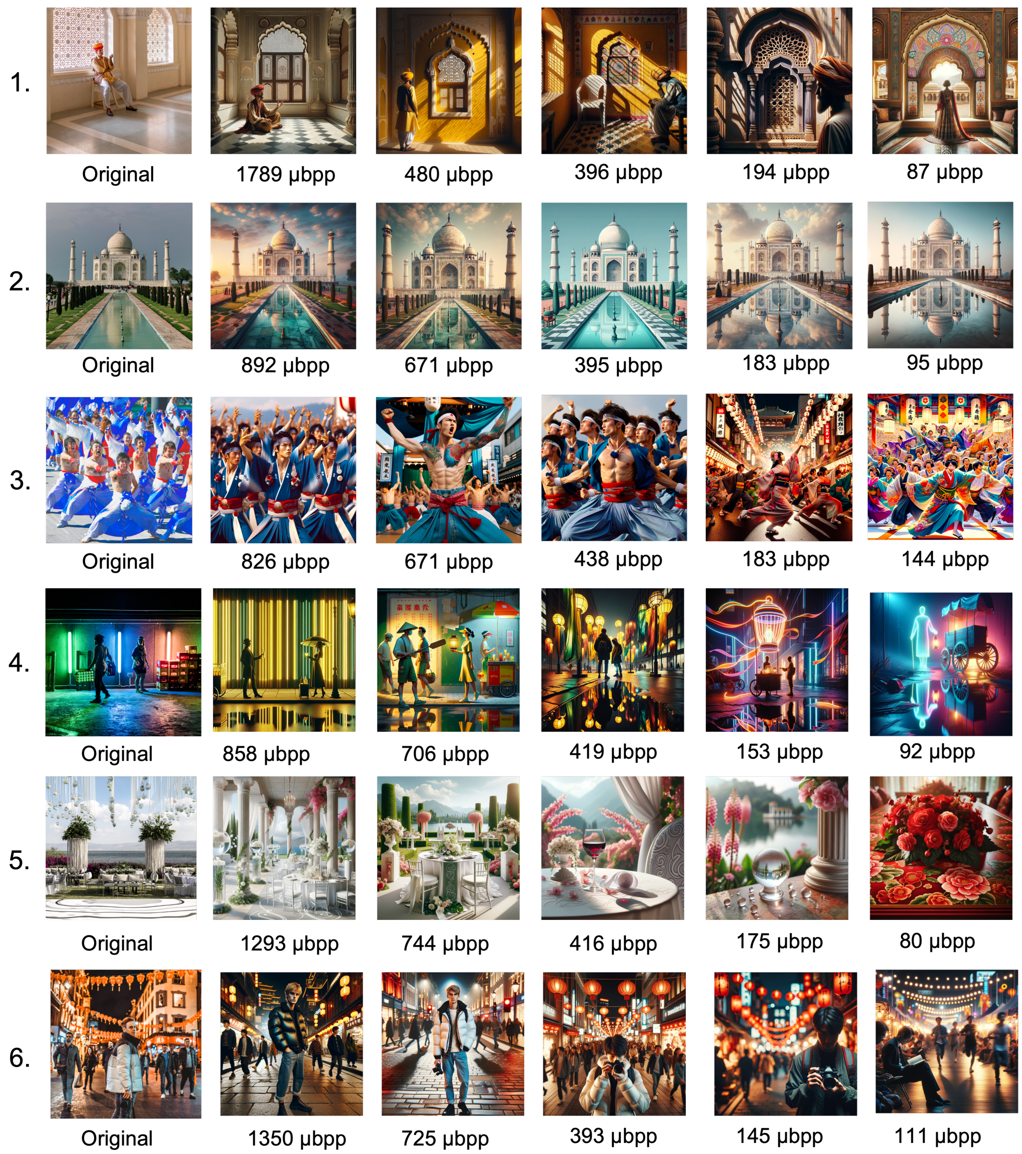

Figure 3: **Compression Examples:** The first example shows the progressive loss of contextual information from tile details, room color, location of the man, sitting vs. standing. The second example, on the other hand, shows that landmarks and proper nouns like the Taj Mahal taken from standard angles can be compressed extremely small to 10s of μbpp since a significant amount of information is captured within a few words. The third example again shows the gradual loss of context, color, gender, and location. The fourth example shows the progressive loss of contextual information including light colors, figure position, and style of the lights. The fifth example shows that with heavy compression it hyper-focuses on certain arbitrary details like the flowers. Finally, the last example shows the loss of information about the jacket color and other details with higher compression.

Figure 4: **More Compression Examples:** These examples show the usefulness of image-specific, variable-rate compression using fewer bits for more common images and gradual decline in quality in most examples at lower bitrates.

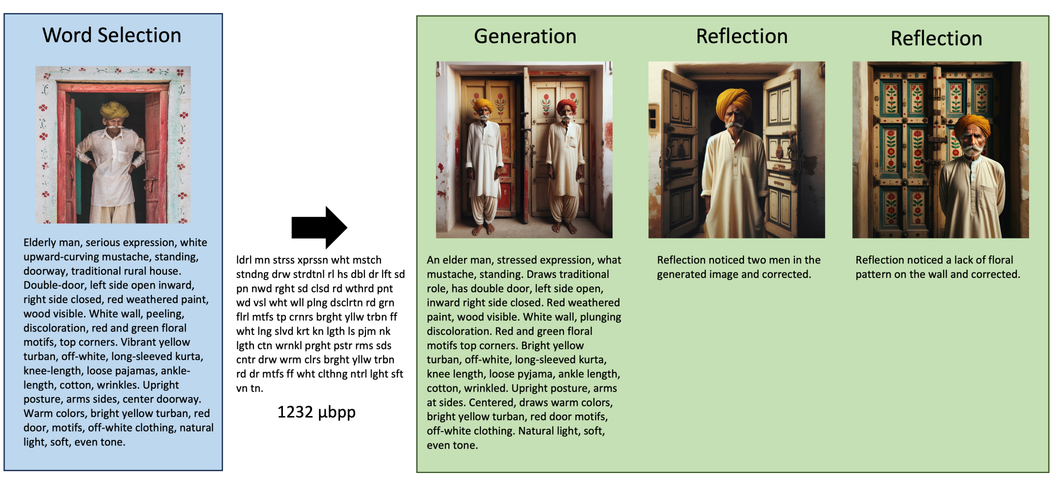

Figure 5: **Bearded Man:** An example of higher bitrates to demonstrate the effectiveness of reflection with sufficient context. Originally, the model produces two men and corrects for its mistake. Then, it has a regression on the floral pattern but identifies it and adjusts appropriately. It follows much of the detail in the uncompressed description.

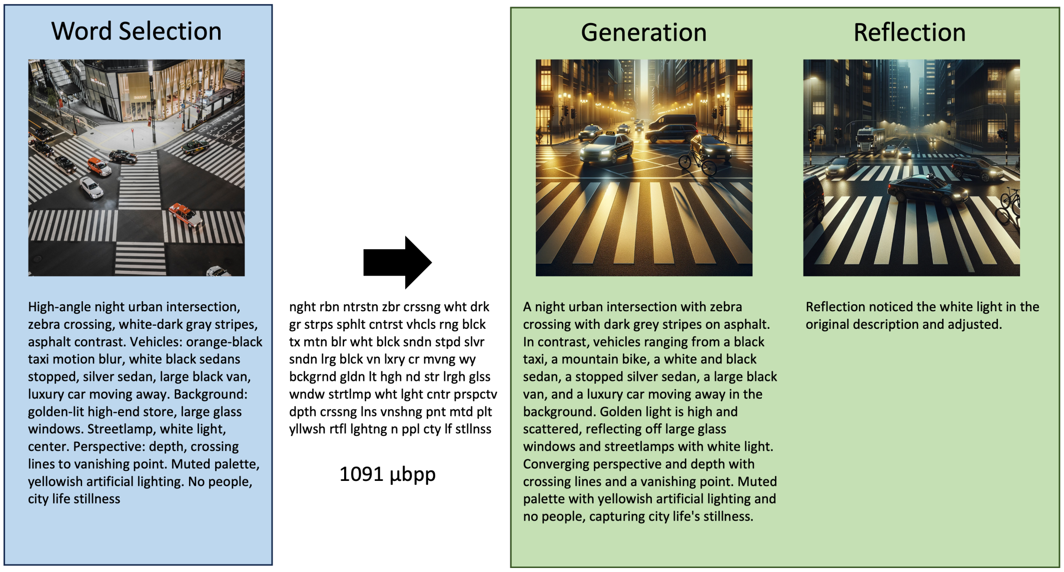

Figure 6: **City Block:** The model was able to recognize the issue with color of the street lighting, but it could not recognize the major difference in orientation. This is a common problem and also occurs with the Brooklyn Bridge example in Figure 2.

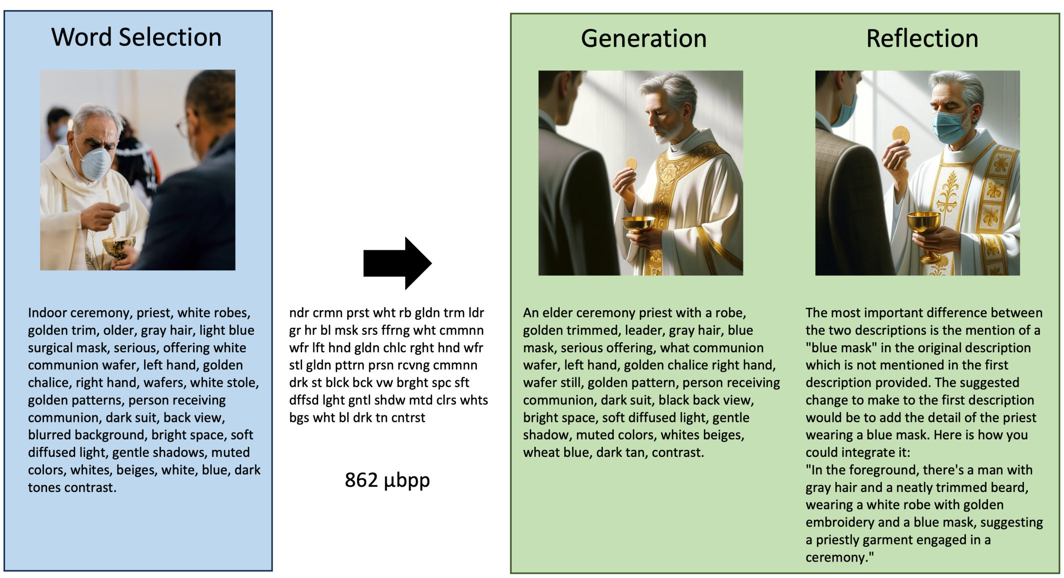

Figure 7: **Priest:** This example demonstrates the ability of DALL-E3 to make in place edits during reflection, although it still errors by adding masks to both the priest and the person receiving communion. It also incorrectly guesses the positions of the priest and other person, since there was no indication in the original description.

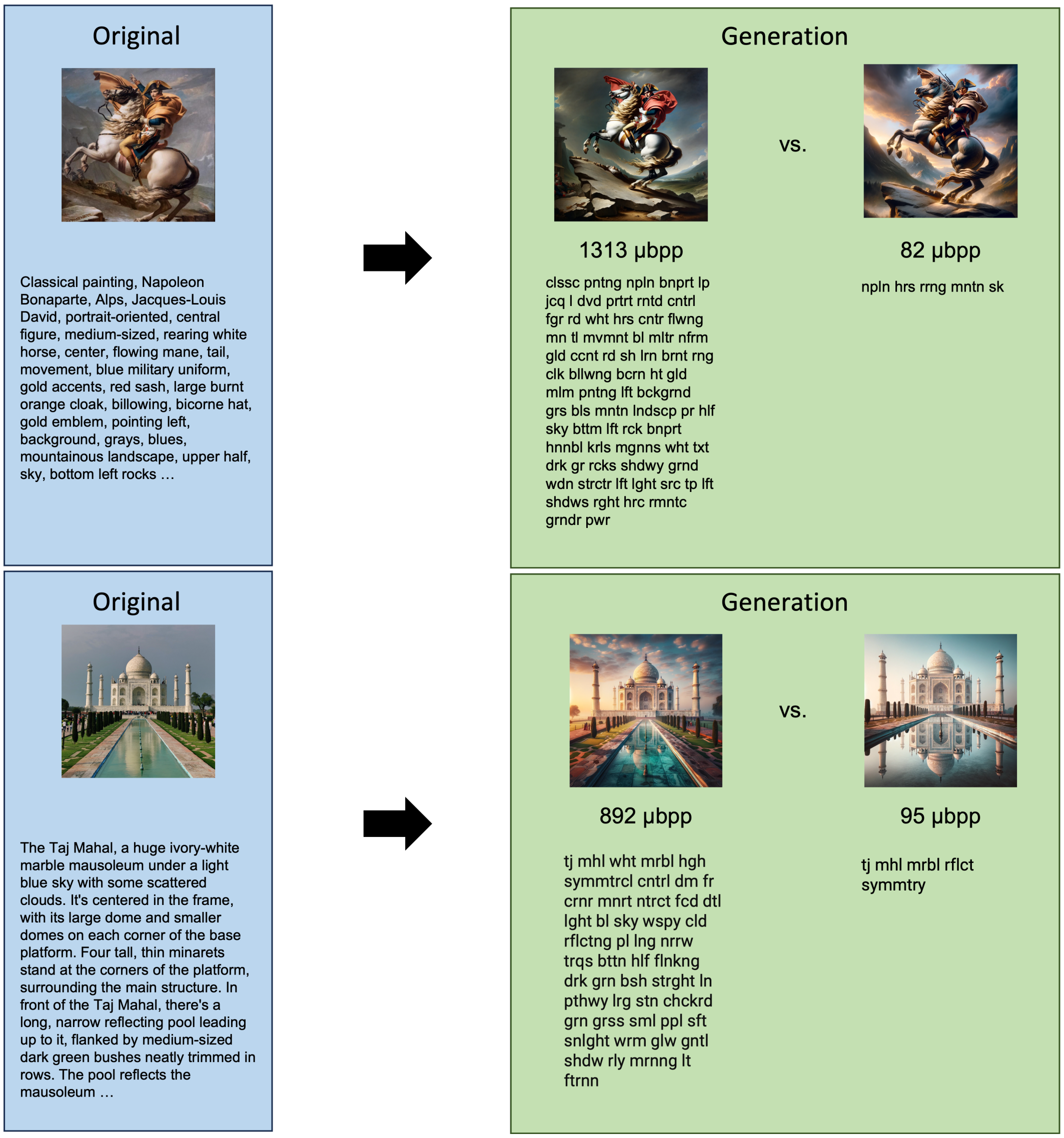

Figure 8: **Common Subjects:** For images of well-known subjects at standard angles, e.g., the Taj Mahal or Napoleon painting, very few words can produce accurate results. There is little appreciable increase in generated image accuracy from longer, more detailed descriptions. This phenomena could potentially be used to improve efficiency through a variable-rate compression algorithm.

B EXPERIMENT DETAILS

All the experiments are run with the GPT4 web interface, which can automatically make calls to the DALL-E3 API. For simplicity, all images are square with a size of 1024×1024, which is a standard output size from DALL-E3. The input images are cropped to this size before passing to GPT4-V. If images are other sizes, then further super-resolution models can be used that can upsample the pixels using neural networks, or likely future decoder models will have better resolution support. Most images are taken from the CLIC (Challenge on Learned Image Compression) dataset, since these images have already been filtered and selected for diversity.

This process is done manually, since at the time of publication DALL-E3 API is still under development and does not support consecutive API calls modifying previously generated images. Without this support, the API cannot support reflection since each call is independent. The encoder and the decoder are opened in different sessions to avoid any shared context. Then, in the first session, the prompts below are used consecutively to describe the image, select the most important words, and then compress the characters in these words. This compressed text is then passed to the second session, which decompresses the text and generates an image based on the description. At the higher bitrates, there is enough context to perform reflection.

During reflection, the same description prompt that is used during encoding is used on the generated image to describe it. Then, the uncompressed description available to the decoder is compared against this new description to select the most salient difference between the images. This difference is passed again to DALL-E3 in the same session, and it can make adjustments to the previous image while attempting to minimize changes elsewhere in the image. In this work, for examples that use reflection, the process continues for a fixed two iterations. This is a hyper-parameter that balances quality and performance, and most images only have a few major potential issues after the initial generation. Yet, in general, it is challenging to create an automated stopping condition for reflection. This is in contrast with using reflection during code generation, where the stopping condition is determined by passing the test cases.

B.1 PROMPTS

Below are the sets of prompts used for all examples in this paper for the encoding, decoding, and reflection tasks. These were unchanged throughout the evaluation, and the current prompts were generated through manual trial-and-error. The strength of the compression was determined by choosing the word count K in the encoder Word Select prompt. Given the inexactness of large language models, this word count is not always honored, yet GPT4-V typically has an error of less than $\pm 10\%$, and this behavior is actually desirable in many cases, since the model often only exceeds the limits with important words.

B.1.1 ENCODE

Describe: Can you describe this photo in as much detail as possible so that someone can recreate it based only on your description? Describe each object and its size in the image with small, medium, large, and huge. Describe the relative locations of all objects from the perspective of the viewer. Describe the colors in each object.

Word Select: This description will be used to regenerate an image. Can you compress this image description to K words with the goal of selecting the most important words that humans would find relevant during the image reconstruction? These should be the most important words. Do not include helper words like prepositions or other unimportant words.

Word Compress: This description will be used to regenerate an image. Please remove all vowels and restrict to the following characters only: n, t, s, r, h, l, d, c, m, f, g, p, b, k, v. No punctuation is allowed. Remove plurals and uppercase letters.

B.1.2 DECODE

Word Decompress: This is a description of an image that has been extremely compressed by removing vowels and punctuation. Keep in mind only these characters were allowed: n, t, s, r, h, l, d, c, m, f, g, p, b, k, v. Please decompress it to its original text.

Generate: Please generate a square image based on this description by following all of the details.

B.1.3 REFLECT

Describe: Can you describe this photo in as much detail as possible so that someone can recreate it based only on your description? Describe each object and its size in the image with small, medium, large, and huge. Describe the relative locations of all objects from the perspective of the viewer. Describe the colors in each object.

Compare: Please compare the image description above to the original description below and highlight the most important difference between the two. Format this difference into a suggested change to make to the description above to make it more like the original description below.

Generate: Please keep the exact same image but make the following change:

C REFLECTION

Reflection is the process of iterative development, which mirrors the human generative process in writing, painting, and other creative tasks. Variants of reflection within language models have been explored in other fields to improve the quality of generative models. For example, Reflexion Shinn et al. (2023) uses it to significantly improve the code generation and achieve state-of-the-art results on the HumanEval task. It first produces candidate code and executes it, analyzes its output or error messages, and then iterates until the code passes a set of test cases or reaches a maximum number of iterations. Our work applies a similar method that generates, analyzes, and iteratively improves images. Specifically, it compares the decoded text description of the original image with a new description of the generated image and highlights the major semantic differences. Then, it suggests the most important change following the prompts in the appendix, and regenerates the image with this change.

Our use of reflection was further motivated by the high descriptive strength of the GPT4-V compared to the generative strength of DALL-E3. This is similar to the classic N=NP problem, which likely suggests it is easier to evaluate that a solution is correct than generate a solution itself. Therefore, the reflection process can make up for relatively poor performance of the decoder through iterative analysis and generation on the initial design. In practice, this process is currently limited by the ability of the decoder to isolate these changes and can lead to regressions during reflection. Sometimes iterations can significantly change the previous image or undo changes made during previous reflection iterations. Overall, however, reflection typically leads to the most serious issues with the generated image being fixed typically within a few iterations.